

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-223214
(43)Date of publication of application : 09.08.2002

(51)Int.Cl. H04L 12/28
G06F 9/46
G06F 13/00
G06F 15/16
H04L 29/08

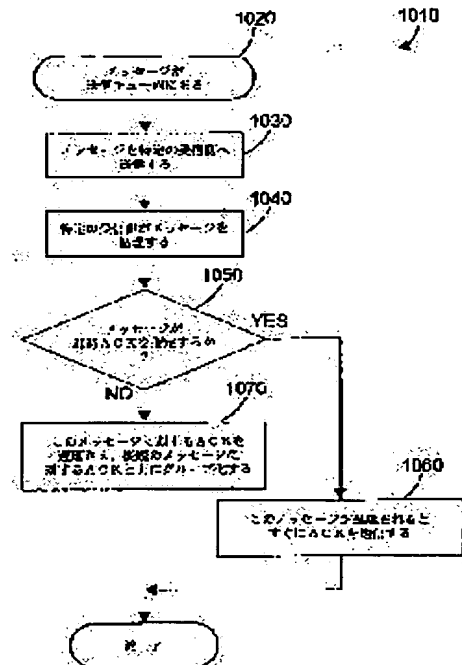
(21)Application number : 2001-336277 (71)Applicant : INTERNATL BUSINESS MACH CORP <IBM>
(22)Date of filing : 01.11.2001 (72)Inventor : BLOCK TIMOTHY R

(30)Priority
Priority number : 2000 718924 Priority date : 22.11.2000 Priority country : US

(54) METHOD AND APPARATUS FOR COMMUNICATION BETWEEN COMPUTER SYSTEMS USING SLIDING TRANSMISSION WINDOW FOR ORDERED MESSAGE IN CLUSTER COMPUTER ENVIRONMENT

(57)Abstract:

PROBLEM TO BE SOLVED: To provide an apparatus and a method by which a sliding transmission window is used for communicating with another computer in a cluster.
SOLUTION: A cluster computer system has a plurality of computer systems (or nodes) which can become members of a group working on a specific task and are connected to each other through one or more networks. Each node includes a cluster communication mechanism incorporating a cluster engine and the sliding transmission window, and one or more service tasks which process messages. The sliding transmission window enables the node to transmit a plurality of messages without waiting for individual acknowledgements to the messages. The sliding transmission window also enables the node to transmit a single confirmation message with respect to a plurality of received messages when the node receives the messages.



LEGAL STATUS

[Date of request for examination] 01.11.2001
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]
[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2002-223214

(P2002-223214A)

(43) 公開日 平成14年8月9日 (2002.8.9)

(51) Int.Cl. ⁷	識別記号	F I	テーマコード [*] (参考)
H 0 4 L 12/28	2 0 7	H 0 4 L 12/28	2 0 7 5 B 0 4 5
G 0 6 F 9/46	3 6 0	G 0 6 F 9/46	3 6 0 F 5 B 0 9 8
13/00	5 2 0	13/00	5 2 0 A 5 K 0 3 3
15/16	6 4 0	15/16	6 4 0 A 5 K 0 3 4
H 0 4 L 29/08		H 0 4 L 13/00	3 0 7 Z
審査請求 有 請求項の数12 O L (全 17 頁)			

(21) 出願番号 特願2001-336277(P2001-336277)

(22) 出願日 平成13年11月1日 (2001.11.1)

(31) 優先権主張番号 09/718924

(32) 優先日 平成12年11月22日 (2000.11.22)

(33) 優先権主張国 米国 (US)

(71) 出願人 390009531

インターナショナル・ビジネス・マシーンズ・コーポレーション

INTERNATIONAL BUSINESS MACHINES CORPORATION

アメリカ合衆国10504、ニューヨーク州アーモンク (番地なし)

(74) 代理人 100086243

弁理士 坂口 博 (外2名)

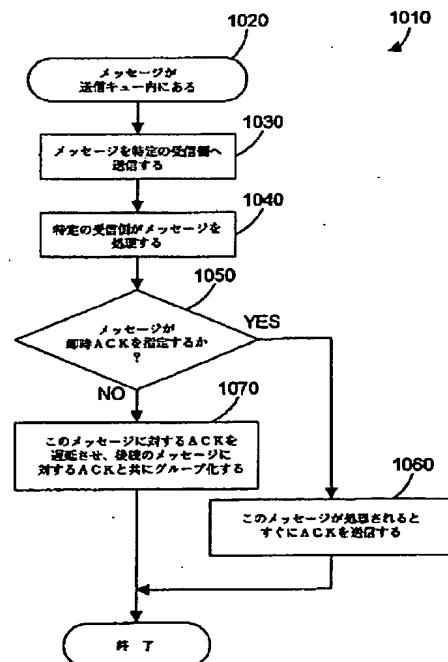
最終頁に続く

(54) 【発明の名称】 クラスタ・コンピュータ環境において順序付きメッセージのためのスライディング送信ウィンドウを用いてコンピュータ・システム間の通信を行う装置および方法

(57) 【要約】

【課題】 クラスタ内の他のコンピュータ・システムと通信するためにスライディング送信ウィンドウを用いる装置および方法を提供する。

【解決手段】 クラスタ・コンピュータ・システムは、特定のタスク上で働くグループのメンバになり得る、1以上のネットワークによって相互に接続された複数のコンピュータ・システム (またはノード) を有する。各ノードは、クラスタ・エンジン、スライディング送信ウィンドウを含むクラスタ通信機構、およびメッセージを処理する1以上のサービス・タスクを含む。スライディング送信ウィンドウは、ノードが、各々のメッセージに対する個々の確認応答を待つことなしに複数のメッセージを発信することを可能にする。スライディング送信ウィンドウは、また、複数のメッセージを受信したノードが、複数の受信メッセージに対して単一の確認メッセージを送信することを可能にする。



【特許請求の範囲】

【請求項1】装置であって、

少なくとも1つのプロセッサ、
前記少なくとも1つのプロセッサに接続されたメモリ、
少なくとも1つの他のコンピュータ・システムに接続されたネットワークへ前記装置を接続するネットワーク・インターフェース、前記メモリに存在し前記少なくとも1つのプロセッサによって実行されるクラスタ通信機構、前記クラスタ通信機構は、次の順序付きメッセージを発信するよりも前に少なくとも1つの他のコンピュータ・システムからの確認メッセージを待つことなしに、前記少なくとも1つの他のコンピュータ・システムへ少なくとも1つの順序付きメッセージを伝達するスライディング送信ウィンドウを有する、
を備える装置。

【請求項2】各順序付きメッセージは、前記順序付きメッセージに対する確認メッセージを遅らせて少なくとも1つの後続の確認メッセージと共にグループ化できるか否か指示する情報を有するヘッダを含む請求項1に記載の装置。

【請求項3】前記確認メッセージは、1から複数の順序付きメッセージを確認する請求項2に記載の装置。

【請求項4】コンピュータ・システムのクラスタを備えるネットワーク・コンピュータ・システムであって、
前記コンピュータ・システムの各々は、
各コンピュータ・システムを、前記クラスタ内の他のコンピュータ・システムへネットワークを介して接続するネットワーク・インターフェース、
メモリ、
前記メモリに存在するクラスタ通信機構、前記クラスタ通信機構は、次の順序付きメッセージを発信するよりも前に少なくとも1つの他のコンピュータ・システムからの確認応答を待つことなしに、前記少なくとも1つの他のコンピュータ・システムへ少なくとも1つの順序付きメッセージを伝達するスライディング送信ウィンドウを有する、
を含むネットワーク・コンピュータ・システム。

【請求項5】各順序付きメッセージは、前記順序付きメッセージに対する確認メッセージを遅らせて少なくとも1つの後続の確認メッセージと共にグループ化できるか否か指示する情報を有するヘッダを含む請求項4に記載のネットワーク・コンピュータ・システム。

【請求項6】クラスタ・コンピュータ環境においてタスクを処理するコンピュータ実装方法であって、前記方法は、
次の順序付きメッセージを発信するよりも前に、順序付きメッセージを受信した前記クラスタ内の各コンピュータ・システムからの確認応答を待つことなしに前記クラスタ内の少なくとも1つの他のコンピュータ・システムへ少なくとも1つの順序付きメッセージを伝達するス

ライディング送信ウィンドウを有する、前記クラスタ内の第1のコンピュータ・システム上で実行するクラスタ通信機構を与えるステップと、
前記クラスタ通信機構が、前記クラスタ内の少なくとも1つの他のコンピュータ・システムへ第1の順序付きメッセージを送信するステップと、
前記クラスタ通信機構が、前記クラスタ内の前記少なくとも1つの他のコンピュータ・システムからの前記第1の順序付きメッセージに対する応答を待つことなしに、
第2の順序付きメッセージを送信するステップとを含む方法。

【請求項7】前記クラスタ内の前記少なくとも1つの他のコンピュータ・システムが、前記第1および第2の順序付きメッセージの双方を確認する単一の確認メッセージを前記クラスタ通信機構へ送信することによって前記第1および第2の順序付きメッセージに応答するステップをさらに含む請求項6に記載の方法。

【請求項8】前記第1および第2の順序付きメッセージは、各々、前記第1および第2の順序付きメッセージに対する確認メッセージを遅らせて少なくとも1つの後続の確認メッセージと共にグループ化できるか否か指示する情報を有するヘッダを含む請求項6に記載の方法。

【請求項9】プログラム・プロダクトであって、

(A) 次の順序付きメッセージを発信するよりも前に、
少なくとも1つの他のコンピュータ・システムからの確認応答を待つことなしにクラスタ内の少なくとも1つの他のコンピュータ・システムへ少なくとも1つの順序付きメッセージを伝達するスライディング送信ウィンドウを有するクラスタ通信機構(A1)を含むコンピュータ・プログラム、

(B) 前記コンピュータ・プログラムを伝達するコンピュータ読み取り可能信号伝達媒体、
を備えるプログラム・プロダクト。

【請求項10】前記信号伝達媒体は、記録可能媒体よりなる請求項9に記載のプログラム・プロダクト。

【請求項11】前記信号伝達媒体は、伝送媒体よりなる請求項9に記載のプログラム・プロダクト。

【請求項12】各順序付きメッセージは、前記順序付きメッセージに対する確認メッセージを遅らせて少なくとも1つの後続の確認メッセージと共にグループ化できるか否か指示する情報を有するヘッダを含む請求項9に記載のプログラム・プロダクト。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、広くデータ処理に関し、特に、ネットワーク上のコンピュータ間のタスクの共用に関する。

【0002】

【従来の技術】コンピュータ時代の始まり以後、コンピュータ・システムは、工学設計、マシンおよびプロセス

・コントロール、そして情報記憶およびアクセスを含む人間の活動の多くの分野において必要不可欠になった。コンピュータの初期において、銀行のような企業、産業界、および政府機関は、彼らの要求を満たす単一のコンピュータを購入していた。しかし、1950年代の初頭までには、多くの企業が複数のコンピュータを有し、1つのコンピュータから他のコンピュータへデータを移す必要性が明白となった。このときから、コンピュータが共働することを可能にするためにコンピュータ・ネットワークが開発され始めた。

【0003】ネットワーク・コンピュータは、どの単一コンピュータも実行できないタスクを実行することができる。加えて、ネットワークは、低コスト・パーソナル・コンピュータ・システムが、より大規模のシステムに接続して、このような低コスト・システムが単独では実行できないタスクを実行することを可能にする。米国内の大部分の企業は、現在、1以上のコンピュータ・ネットワークを有する。ネットワークのトポロジおよびサイズは、ネットワーク化されるコンピュータ・システムおよびシステム管理者の設計に従って変化し得る。実際に、企業が複数のコンピュータ・ネットワークを有することは非常に一般的なことである。多くの巨大企業は、企業内の大多数のコンピュータを相互に効果的に接続するローカル・エリア・ネットワーク（LAN）および広域ネットワーク（WAN）の複雑な混合を有する。

【0004】ネットワーク上で共につながれた複数のコンピュータに関して、ネットワーク・コンピュータを用いて、タスクの異なる部分をネットワーク上の異なるコンピュータに委託し、それらが続いて各自の個々の部分を並列に処理できることによってタスクを完了することができるということが速やかに明白となった。ネットワーク上の共用コンピューティングに対する1つの具体的な構成において、タスクの異なる部分に対して並列に働くことができるネットワーク上のコンピュータ・システムのグループを定義するためにコンピュータ“クラスタ”の概念が用いられてきた。加えて、コンピュータ・クラスタは、きわめて信頼性が高いサービスを伴う“単一システム・イメージ”を提供する。クラスタ内の複数のシステムは、ユーザにとって1つのコンピュータ・システムのように見え得る、そして、ユーザが必要とするサービスおよびリソースは、たとえクラスタ内のコンピュータ・システムのうちの1つが故障し、あるいはメンテナンスのためにダウンされていても常に利用できる。

【0005】クラスタ内のコンピュータが共働してタスクを実行するための1つの方法は、順序付きメッセージの概念を用いる。順序付きメッセージ・システムにおいて、各々のメッセージは、典型的にIPマルチキャストを用いて全てのノードへ伝達され、メッセージの順序は、全てのノードが同じ順序で一定の発信元からのメッセージを認識するように強制される。先行技術クラスタ

・コンピュータ環境において、各々のメッセージは、続けて次のメッセージを処理するよりも前に各ノードによって処理される。すなわち、順序付きメッセージを使用するクラスタ・コンピュータ環境において通信を行う先行技術は、単一メッセージを備える送信ウィンドウを有し、これは、1つの固定された送信ウィンドウ・サイズである。

【0006】

【発明が解決しようとする課題】“スライディング送信ウィンドウ”の概念は、伝送制御プロトコル（TCP）二地点間メッセージと関連して技術上既知である。スライディング送信ウィンドウは、複数のメッセージが、次のメッセージを送信するよりも前に各々のメッセージに対する個々の確認応答を待つことなしに送信されることを可能にする。スライディング送信ウィンドウは、TCPを用いる二地点間通信として知られているが、スライディング送信ウィンドウは、それが現在まで解決されていない特定の問題を生じさせるので、クラスタ・コンピュータ環境において使用されなかった。特に、IPはメッセージの順序を強制していないので、個々のノードに対するマルチキャストである順序付きメッセージを、全てのノード上で同じ順序で処理するという要件は、先行技術TCPスライディング送信ウィンドウを用いて実現できない。こういう理由で、クラスタ・コンピュータ環境におけるIPマルチキャスト通信は、スライディング送信ウィンドウの使用によって利益を得られなかった。クラスタ・コンピュータ環境において使用可能なスライディング送信ウィンドウを提供するメカニズムおよび方法がなければ、クラスタ・コンピュータ・システムの性能は、現行の1つの固定された送信ウィンドウ・サイズによって制限され続けることになる。

【0007】

【課題を解決するための手段】好適な実施の形態によれば、クラスタ・コンピュータ・システムは、特定のタスク上で働くグループのメンバになり得る、1以上のネットワークによって相互に接続された複数のコンピュータ・システム（またはノード）を含む。各々のノードは、クラスタ・エンジン、スライディング送信ウィンドウを有するクラスタ通信機構、およびメッセージを処理する1以上のサービス・タスクを含む。スライディング送信ウィンドウは、ノードが、各メッセージに対する個々の確認応答を待つことなしに複数のメッセージを発信することを可能にする。スライディング送信ウィンドウは、また、複数のメッセージを受信したノードが、複数の受信メッセージに対して単一の確認メッセージを送信することを可能にする。クラスタ内の他のコンピュータ・システムと通信するためにスライディング送信ウィンドウを用いることにより、クラスタ内の通信トラフィックは著しく削減され、それによってクラスタの全体性能を高める。加えて、同時に送信された複数のメッセージ間の

待ち時間は、非常に効果的に短縮される。

【0008】本発明の上述および他の特徴および利点は、添付の図面において例示される下記の本発明の好適な実施の形態のより詳細な説明から明らかにすることができる。

【0009】

【発明の実施の形態】本発明の好適な実施の形態を、同一の指示が同一の要素を示す添付の図面と関連して以下に説明する。

【0010】本発明は、ネットワーク上で接続されたコンピュータ上でタスクの一部を共用することによって達成される。ネットワーク概念に精通していない読者のために、後述の簡潔な概要は、読者が本発明を理解するのに役立ち得る背景情報を提供する。

【0011】1. 概要

ネットワーク・コンピュータ・システム

コンピュータをネットワーク上で相互に接続することは、ある種のネットワーキング・ソフトウェアを必要とする。長年にわたって、ネットワーキング・ソフトウェアの能力および複雑さは著しく増大した。ネットワーキング・ソフトウェアは、典型的に、ネットワーク上のコンピュータ間で情報を交換するためのプロトコルを定義する。多くの異なるネットワーク・プロトコルが技術上知られている。商業的に入手可能なネットワーキング・ソフトウェアの代表例は、Novell NetwareおよびWindows(R) NTであり、各々がコンピュータ間で情報を交換するための異なるプロトコルを実装する。

【0012】近年非常にポピュラーになった1つの重要なコンピュータ・ネットワークは、インターネットである。インターネットは、コンピュータおよびネットワークの急増から発展し、コンピュータ・システムの複雑なワールドワイド・ネットワークへ進化した。インターネットを用いて、ユーザは、単一のワークステーションから世界中のコンピュータにアクセスできる。TCP/IP(伝送制御プロトコル/インターネット・プロトコル)は、インターネットを介した2つのコンピュータ間の二地点間通信のために今日広く使われているネットワーク・プロトコルの代表例である。加えて、TCP/IPの使用は、また、企業内部のさらに多くのローカル・エリア・ネットワーク(LAN)およびイントラネットに迅速に広がっている。

【0013】ユーザ・データグラム・プロトコル(UDP)は、既知のネットワーク・プロトコルの他の例である。UDPは、TCPと関連したオーバーヘッドを多くは備えないだけでなく、TCPの信頼性を備えない。TCPにおいて、2つのコンピュータ・システムは、当該2者の間に“コネクション”を設定することによって二地点間で通信する。受信側のノードが、送信側のノードによって送信されたメッセージを受信することができな

い場合には、送信側のノードは、受信側のノードがメッセージを確認しなかったということを認識し、メッセージを再送することになる。他方、UDPは、“コネクション”を扱わず、メッセージの受信を確認するための設計された方法を備えない。その結果、送信側のコンピュータ・システムは、メッセージが受信されたか否かを識別する方法を備えない。UDPは、コンピュータ・クラスタ内のIPマルチキャスト環境において首尾よく用いられてきた、しかし、重要なシステム・レベル・コードに、UDPによって送信され受信されたメッセージを管理して信頼性が高い通信を保証することを要求する。要するに、UDPを用いることによってTCPのオーバーヘッドの一部を除去することにより、より低いレベルの実装がシステム・レベル・コードのプログラマにとって利用可能となり、それによってコンピュータ・クラスタのための低レベル通信プロトコルの実装に、より高い柔軟性を与える。

【0014】コンピュータ・クラスタ

先行技術は、コンピュータ・システムのグループに問題の異なる部分上で働かせることの利点を認めている。コンピュータの“クラスタ”の概念は、発展して、より大きいタスクの一部を共用できるネットワーク・コンピュータの事前定義されたグループを包含した。クラスタの1つの具体的な実装は、クラスタ内のコンピュータ間で通信を行うために順序付きメッセージを用いる。順序付きメッセージ・システムにおいて、各メッセージは、全てのノードに伝達され、メッセージの順序は、全てのノードが同じ順序でメッセージを認識するように強制される。複数のコンピュータへ順序付きメッセージを同時にブロードキャストする1つの既知の方法は、IPマルチキャストを使用する。

【0015】図1を参照すると、5つのコンピュータ・システム(またはノード)110からなる単純クラスタ100が示される。これらのノード間の接続は、論理接続を表し、物理接続は、クラスタ内のノードが相互に論理的に通信可能である限りは好適な実施の形態の範囲内で変化する。クラスタ内部で、共働してタスクを成しとげるノードの論理グループに相当する1以上の“グループ”が定義可能である。グループ内の各ノードは、そのグループの“メンバ”といわれる。図2に示されるように、先行技術クラスタ内の各ノード210は、システム・レベル・コード290およびカーネル292を有するオペレーティング・システムを含む。

【0016】カーネル292は、コンピュータ・システム・ハードウェアと直接対話する低レベル・オペレーティング・システム・コードを表す。一番下の層は、IP/物理層280であり、これは物理的な通信媒体を通じて通信するオペレーティング・システム・ソフトウェアの層である。IP/物理層280の上にUDP層270があり、これはコンピュータ・システム間でメッセージ

を交換するためのネットワーク・プロトコルを提供する。クラスタ・トポロジ・サービス262およびクラスタ通信260は、UDP層270の上に存在する。クラスタ・トポロジ・サービス262は、クラスタの現行トポロジ・ビューを保持し、必要に応じてクラスタにメンバを追加し、あるいはクラスタからメンバを削除することによってクラスタのトポロジを変更するためのサポートを提供する。クラスタ通信260は、クラスタ内の各コンピュータ・システムからの順序付きメッセージの伝送および受信のサポートを提供する機構である。クラスタ通信機構260は、単一ソースへのおよび単一ソースからのメッセージの順序を保証する。しかし、それぞれに異なるコンピュータ・システムへのあるいはそれぞれに異なるコンピュータ・システムからのメッセージ間の順序は保証しない。クラスタ・エンジン250（CLUEとしても知られている）は、クラスタ通信機構260によって他のノードからメッセージを受信し、全てのソースからの全てのメッセージの総合的な順序を保証する。CLUE250は、クラスタ内のノード間の順序付きメッセージを強制するソフトウェア・プロセスである。CLUE250が、グループに向けられたそのメンバからのメッセージを受信するとき、CLUE250は、利用可能である場合にIPマルチキャストを典型的に使用するクラスタ通信機構を介して、グループの全ての登録済みメンバに対してメッセージを送信する。CLUEコードの一部は厳密にはカーネル292の一部とみなされ、一方CLUEの他の部分は厳密にはシステム・レベル・コード290とみなされ、これが、クラスタ・エンジン250が各々の一部を含むということが図2に示される理由であるということを留意されたい。

【0017】クラスタ制御層240およびクラスタ・グループ・マネージャ242が、CLUE層250の上に位置する。クラスタ制御240は、ノードに対するクラスタ化の構成および活動化を管理し、典型的に、クラスタ環境の管理に適切である様々なクラスタ初期化およびノード管理オペレーションをサポートする。クラスタ・グループ・マネージャ242は、全クラスタのグループ・メンバーシップ状況情報のコピーを同期的に保持する一方、クラスタ・ライブラリ関数230は、クラスタに他のサポート・サービスを提供する。クラスタ化APIコンポーネント220は、ジョブ/アプリケーション214（例えば図2に示されるジョブ/app214Aおよびジョブ/app214B）を介して基礎クラスタ化機能に外部インターフェースを提供する。クラスタ・マネージャ212は、それによってユーザがクラスタ通信パラメータの変更を開始できるユーザ・インターフェースを提供する。

【0018】クラスタ通信機構260の先行技術実施例が図3に示される。サイズが1つの固定送信ウィンドウ310は、送信されるべきメッセージを含む送信キュー

320、どのメッセージが現在作動されているか指示する現行メッセージ属性330およびどのノードが現行メッセージを確認したかに関する情報を含むACKインジケータ340と連係して使用される。

【0019】2. 詳細な説明

本発明の好適な実施の形態による装置および方法は、クラスタ・コンピュータ環境におけるスライディング送信ウィンドウを提供する。スライディング送信ウィンドウは、複数のメッセージが、各メッセージに対する個々の応答を待つことなしに送信されることを可能にする。その代わりに、複数のメッセージの受信を確認する単一の確認メッセージを送信可能である。クラスタ内のコンピュータ・システム間の通信のためのスライディング送信ウィンドウの使用は、ネットワーク・トラフィックに大きな削減をもたらし、それによってクラスタの性能を高める。

【0020】図4を参照すると、好適な実施の形態に従うクラスタ通信機構460は、グループの全てのメンバによる各メッセージの受信を個別に確認することなしに、クラスタ内のグループのメンバに複数のメッセージを送信することを可能にするスライディング送信ウィンドウ410を含む。送信キュー320は、好ましくは、図3の先行技術実施例におけるのと同じであり、しかしながら、代替の実施もまた可能である。現行メッセージ・キュー430および保留ACKキュー440は、先行技術におけるそれらの同等物330および340と個々に比べて機能が拡張され（すなわち、単一のデータ属性からキューあるいはベクトル（2次元の）・エンティティへ）、複数のメッセージを処理する。

【0021】図4のスライディング送信ウィンドウ410の例が図5において図示される。この例に関して、図5の510において示されるように、送信ウィンドウは最初にゼロである（メッセージがない）ということが仮定される。5つのメッセージm1～m5は、520において示されるように、これらのメッセージのうちのどれに対してもACKを受信することなしに、順々に発信されるということが仮定される。スライディング送信ウィンドウは、520において示されるように、それが幅が5つのメッセージm1～m5になるまで、一度に一つのメッセージを拡大する。次に、m1およびm2に対するACKが受信され、これが、530において示されるように、スライディング送信ウィンドウを3つのメッセージm3～m5の幅へ縮小させるとということが仮定される。2つのそれ以外のメッセージm6およびm7が続いて送信され、これが、スライディング送信ウィンドウのサイズを5つのメッセージm3～m7へ拡大させるとということが仮定される（540）。続いて、これらのメッセージの全てに対するACKが受信され、これが、550において示されるように、スライディング送信ウィンドウをゼロへスライドさせ、560において示されるよ

うにゼロのサイズへ縮小させるということが假定される。図5は、なぜ送信ウィンドウが“スライディング”送信ウィンドウと称されるか説明する。そのサイズは、保留中のメッセージの総数および、確認されたメッセージの数に従って変化する（あるいは上下する）。

【0022】図6を参照すると、コンピュータ・システム600は、拡張IBM iシリーズ・コンピュータ・システムであり、好適な実施の形態に従って相互にネットワーク化可能な1つの適切なタイプのノード110

(図1)を表す。当業者は、本発明のメカニズムおよび装置が、他のコンピュータ・システムと共にネットワーク化可能ないかなるコンピュータ・システムに対しても等しく適用されるということを理解できる。図6に示されるように、コンピュータ・システム600は、主記憶装置620、大容量記憶インターフェース630、端末インターフェース640およびネットワーク・インターフェース650に接続されたプロセッサ610を含む。これらのシステム・コンポーネントは、システム・バス660の使用を通して相互接続される。大容量記憶インターフェース630は、大容量記憶装置（例えば、直接アクセス記憶装置655）をコンピュータ・システム600へ接続するために使用される。直接アクセス記憶装置655の1つの具体的な例は、フロッピー（R）・ディスク695へデータを格納し、フロッピー（R）・ディスク695からデータを読み取ることができるフロッピー（R）・ディスク・ドライブである。

【0023】主記憶装置620は、データ622およびオペレーティング・システム624を含む。データ622は、コンピュータ・システム600内のあらゆるプログラムへの入力あるいはあらゆるプログラムからの出力に使えるあらゆるデータを表す。オペレーティング・システム624は、OS/400として産業界で知られているマルチタスク・オペレーティング・システムである。しかしながら、当業者は、本発明の趣旨および範囲は、どの1つのオペレーティング・システムにも限定されないということを理解できる。オペレーティング・システム624は、OSシステム・レベル・コード690およびカーネル692を含む。システム・レベル・コード690は、図2におけるOSシステム・レベル・コード290と同一または同種であってもよく、あるいは好適な実施の形態の範囲内で完全に異なってもよいということを留意されたい。OSカーネル692は、クラスタ内の他のノードと通信するために使用されるスライディング送信ウィンドウ410を含むクラスタ通信機構460を有する。OSカーネル692は、IPマルチキャストによってグループの他のメンバと通信するためにクラスタ通信機構460によって使用されるIP/物理層（図2の280と同様）の一部であるIPマルチキャスト・サポート626を付加的に含む。好適な実施の形態は、あらゆる組み合わせにおいてコンピュータ・ネット

ワーク上の二地点間通信およびマルチキャスト通信の双方に明らかに及ぶということを留意されたい。

【0024】コンピュータ・システム600は、コンピュータ・システム600のプログラムが、まるでそれらが複数のより小容量のストレージ・エンティティ例えば主記憶装置620およびDASDデバイス655に対するアクセスの代わりに、大容量の単一のストレージ・エンティティに対するアクセスのみを有するかのよう機能することを可能にする、周知の仮想アドレッシング機構を使用する。したがって、データ622およびオペレーティング・システム624は、主記憶装置620に存在するように示されているが、当業者は、これらのアイテムが、必ずしも同時に主記憶装置620に全て完全に含まれる必要はないということを理解できる。用語“メモリ”は、コンピュータ・システム600の全体の仮想メモリを一般的に指示するためにこの中で使用されることがも留意されたい。

【0025】プロセッサ610は、1以上のマイクロプロセッサおよび/または集積回路から構成可能である。プロセッサ610は、主記憶装置620に格納されたプログラム命令を実行する。主記憶装置620は、プロセッサ610がアクセスできるプログラムおよびデータを格納する。コンピュータ・システム600が始動するとき、プロセッサ610は、オペレーティング・システム624を構成するプログラム命令を最初に実行する。オペレーティング・システム624は、コンピュータ・システム600のリソースを管理する高性能のプログラムである。これらのリソースの一部は、プロセッサ610、主記憶装置620、大容量記憶インターフェース630、端末インターフェース640、ネットワーク・インターフェース650およびシステム・バス660である。単一のプロセッサおよび単一のシステム・バスのみを含むコンピュータ・システム600を示したが、当業者は、複数のプロセッサおよび/または複数のバスを有するコンピュータ・システムを用いて本発明を実施可能であるということを理解できる。

【0026】端末インターフェース640は、1以上の端末装置665をコンピュータ・システム600へ直接接続するために使用される。非インテリジェント（すなわち、ダム）端末装置あるいは完全プログラマブル・ワークステーションであり得るこれらの端末装置665が使用され、システム・アドミニストレータおよびユーザがコンピュータ・システム600と通信することを可能にする。しかしながら、端末インターフェース640が1以上の端末装置665との通信をサポートするために与えられる場合には、ユーザとの全ての必要な対話および他のプロセスは、端末インターフェース640によって発生可能であるので、コンピュータ・システム600は、必ずしも端末装置665を必要としないということを留意されたい。

【0027】ネットワーク・インターフェース650は、他のコンピュータ・システムおよび/またはワークステーション（例えば図6の675）をネットワーク670を越えてコンピュータ・システム600へ接続するために使用される。ネットワーク670は、ネットワーク670上のコンピュータ・システム600および他のコンピュータ・システム間の論理接続を表す。本発明は、ネットワーク接続670が今日のアナログおよび/またはデジタル手法を用いて成されるか、それとも将来のネットワーク・メカニズムによって成されるかわからず、たとえどのようにしてコンピュータ・システム600が他のコンピュータ・システムおよび/またはワークステーションと接続されていても等しく適用され得る。加えて、多くの異なるネットワーク・プロトコルがネットワークを実装するために使用可能である。これらのプロトコルは、コンピュータがネットワーク670を越えて通信することを可能にする特殊コンピュータ・プログラムである。TCP（伝送制御プロトコル）は適切なネットワーク・プロトコルの一例である。

【0028】ここで、完全に機能的なコンピュータ・システムの状況において本発明を説明してきた、そして説明し続けることになるが、当業者は、様々な形式のプログラム・プロダクトとして本発明を配布できるということ、および、配布を実際に実行するために使用される信号伝達媒体の特定のタイプにかかわらず本発明を等しく適用できるということを理解し得るということを留意されたい。適切な信号伝達媒体の例は、記録可能なタイプの媒体、例えばフロッピー（R）・ディスク（例えば図6の695）およびCD ROM、そして伝送タイプの媒体例えばデジタルおよびアナログ通信リンクを含む。

【0029】図7を参照すると、3つのノード600A（ノードA）、600B（ノードB）および600C（ノードC）が全て、ローカル・エリア・ネットワーク（LAN）上で相互に接続されているネットワーク構成例700が示される。これは、技術上知られているコンピュータ・クラスタの最も一般的なネットワーク構成である。図8は、図7のネットワークに対する先行技術のもとでのネットワーク・トラフィックを説明する。ノードA内の送信キュー420は、それらが送信キュー420へ書き込まれたのと同じ順序でノードBおよびCへ送信される必要がある3つの順序付きメッセージを有するということが仮定される。m1が送信キュー420に最初に受信され、m2およびm3がそれに続くということが仮定される。まず、ステップ810において、ノードAがノードBにm1を伝達する。次に、ステップ812において、ノードAがノードCにm1を伝達する。ノードAは、m2を送信する前に、確認メッセージ（ここではACKと称される）がm1を受信した各ノードから受信されるまで待機しなければならない。それゆえ、ノード

ノードAは、ステップ820においてm1に対するACKがノードBから受信され、ステップ822においてm1に対するACKがノードCから受信されるまで待機する。グループの全ての他のメンバが対応するACKを用いてm1に回答したので、ノードAは、ノードB（ステップ830）およびノードC（ステップ832）に対してm2を発信可能である。ノードAは、ノードB（ステップ840）およびノードC（ステップ842）の双方からACKが受信されるまで再び待機しなければならない。いったんm2に対する全てのACKが受信されると、ノードAは、ノードB（ステップ850）およびノードC（ステップ852）に対してm3を発信できる。ノードAは、ノードB（ステップ860）およびノードC（ステップ862）からACKが受信されるまで再び待機する。図8は、ノード（例えばノードA）は、次のメッセージを発信するよりも前にグループの各メンバからのACKを待たなければならないということを図示する。これは、メッセージの処理が、受信されたのと同じ順序で実行されるということを保証するために行われる。しかしながら、次のメッセージを発信するよりも前に各メッセージに対するACKを待つことは、全ての発信メッセージを直列化することによる障害をもたらす。発信メッセージのこの直列化は、次のメッセージへ移るよりも前に各ACKを待つことによるシステム性能の不利をもたらす。

【0030】図7におけるのと同じネットワーク構成に関して図9を概説することにより、好適な実施の形態の概念を図7および8の例と容易に対比し対照することができる。好適な実施の形態において、ノードが、各個別のメッセージに対する確認信号を待つことなしに複数の順序付きメッセージを発信することを可能にするスライディング送信ウィンドウが用いられる。したがって、図9のノードAは、ステップ910においてノードBに対してm1を発信し、ステップ912においてノードCに対してm1を発信する。ノードAは、続いて、ノードBおよびCからのm1に対するACKを待つことなしに、ステップ920および922においてm2を、ステップ930および932においてm3を発信することができる。ノードBおよびCの各々は、続いて、図8の先行技術において示されるようなノードBからの3つの別個のACKおよびノードCからの3つの別個のACKを必要としないで、メッセージm1、m2およびm3の全てを同時に確認する単一のACKを送信することができる。好適な実施の形態の利点は、従って2倍である。第1に、ノードAは、各メッセージに対する各ノードからの個別のACKを待つことなしにメッセージを発信し続け、それによってクラスタに対するワーク・パイプラインをより完全に保つことができる。第2に、m1、m2およびm3を受信したノードは、複数のメッセージを同時に確認する単一のACKを用いて応答し、それによ

てクラスタ・コンピュータ環境において必要とされるACKの数を著しく削減することができる。送信側のノードが、次のメッセージを発信するよりも前に各メッセージに対する個別の確認応答を待つことなしに複数のメッセージを発信することを可能にすることにより、そして受信側が、単一の確認応答を用いて複数のメッセージを確認することを可能にすることにより、好適な実施の形態に従うクラスタ・コンピュータ・システムの性能が著しく増大する。

【0031】図9の通信は、ノードAとノードBおよびCとの間の二地点間通信として示されるということを留意されたい。しかしながら、IPマルチキャストを用いてノード間で通信を行うことは、同様に、好適な実施の形態の範囲内である。この図において、図9のステップ910および912は、IPマルチキャストを用いてBおよびCの双方に対してm1をブロードキャストする単一のステップに併合可能である。同様に、ステップ920および922は、単一のIPマルチキャスト・ステップに置き換え可能であり、ステップ930および932は、単一のIPマルチキャスト・ステップに置き換え可能である。好適な実施の形態は、二地点間通信、マルチキャスト通信、およびこれら2つのあらゆる適切な組み合わせに明らかに及ぶ。

【0032】図10を参照すると、方法1010は、メッセージがノードの送信キュー内にある（ステップ1020）ときに、好適な実施の形態に従う1つの例示的な方法において実行されるステップを示す。メッセージが、特定の受信側へ送信される（ステップ1030）。メッセージは、グループ内の全てのノードに対するマルチキャスト・メッセージであってもよく、あるいは、各受信側ノードに直接伝達される二地点間メッセージであってもよい。受信側のノードは、グループ内のノードと別個であってもよく、特定のソースからのメッセージの順序を保存しながら、マルチキャストおよび二地点間通信を混合することを可能とするということを留意されたい。特定の受信側は、続いて、メッセージを処理する（ステップ1040）。メッセージが即時応答（またはACK）を指定する場合（ステップ1050=YES）には、受信側がメッセージを処理するとすぐに各受信側によってACKが送信される（ステップ1060）。一方、メッセージが即時応答を指定しない場合（ステップ1050=NO）には、メッセージに対するACKは遅延され、後続のメッセージに対する1以上のACKと共にグループ化される（ステップ1070）。このように、受信側は、ACKを一緒にして、複数のメッセージが確認されているということを指定する単一のACKにグループ化することができる。

【0033】他のノードに対するブロードキャストである各メッセージは、図11において例として示されるヘッダ1100のように、様々な情報を有するヘッダを含

む。ヘッダ1100は、メッセージを送信したクラスタ通信機構のバージョン番号を識別するバージョン・フィールド1110、メッセージのタイプを識別するタイプ・フィールド1112、メッセージに関する情報を提供する様々なフラグを有するフラグ・フィールド1114、およびメッセージの長さを識別する長さフィールド1116を含む。ソースIDフィールド1120は、どのノードがメッセージを送信したか識別し、一方、宛先IDフィールド1130は、どのノードがメッセージを受信すべきかを識別する。ソースIPフィールド1140は、送信側のインターネット・プロトコル（IP）・アドレスを指定し、一方、宛先IPフィールド1150は、宛先ノードのIPアドレスを指定する。コネクション番号フィールド1160は、2つのノード間あるいはノードとサブネット（例えばそのサブネット上のノードのグループ）間のコネクションに対応する番号を含む。シーケンス番号フィールド1170および1180は、送信された特定のメッセージに対するシーケンス番号を指示する順次的な番号を指示する。ネクスト・フィールド1190は、seq1フィールド1170内の値を反映する現在は未使用のフィールドである。

【0034】図11のフラグ・フィールド1114は、図12に示される遅延ACKフラグ1200を含む。遅延ACKフラグは、ACKメッセージが即時に送信される必要があるか否かを指示するために使用される。遅延ACKフラグがセットされている場合には、受信側は待機し、後の時点においてこのメッセージを含む個々のメッセージに対するACKを合わせてグループ化することができる。遅延ACKフラグがクリアされている場合には、受信側は、即時に確認応答を行う必要がある。即時に確認応答を行うとは、メッセージが処理された後にACKを送信することを単に意味し、このACKは、つまり、このメッセージに加えてあらかじめ処理された1以上のメッセージに対するグループACKであり得るということを留意されたい。遅延ACKフラグの意義は、ACKがこのメッセージのみに対する単一のACKであるにせよ、あるいはこのメッセージおよび1以上のより早期のメッセージに対するACKであるにせよ、送信側が次のメッセージを発信可能となる前に、このメッセージに対するACKが送信側によって必要とされるということである。

【0035】図13を参照すると、ネットワーク構成例が、好適な実施の形態の概念をさらに説明するために示される。この構成において、ノードAはLAN1上のノードであり、ノードBおよびDはLAN2上のノードであり、ノードCはLAN3上のノードである。LAN1、LAN2およびLAN3は、全て広域ネットワークWAN1によって互いに接続されている。図13のネットワーク構成は、図7の単純LAN構成よりも複雑であり、以下に説明される好適な実施の形態の顕著な特色の

一部を説明するのに役立つ。

【0036】図14は、図13に示されるノードAの機能の一部を説明する。ノードAは、送信キュー420、LAN2に対するシーケンス番号を追跡するオブジェクト1410、およびLAN3に対するシーケンス番号を追跡するオブジェクト1440を含む。送信キュー420は、4つのメッセージm1~m4を含む。この例に関して、m1、m2およびm4は、ノードA、B、CおよびDであると定義されるグループXに基づくメッセージであるということが仮定される。m3は、ノードBに対する二地点間メッセージであるということも仮定される。図13のノードA、B、CおよびD間の対話は図15に示される。

【0037】図15の詳細を説明する前に、図15の基礎にある概念を説明する必要がある。ノードAは、メッセージが紛失し再送が必要であるとみなされるよりも前にACKを受信することが許容される最大時間にセットされているメッセージ・タイマを含むということが仮定される。ノードAは、また、最終メッセージ宛先レジスタを含み、最終メッセージ宛先を、現行メッセージ宛先あるいは次のメッセージ宛先と比較することができる。ノードB、CおよびDは、各々、遅延ACKタイマを含むということも仮定される。各ノードの遅延ACKタイマは、必要以上に多くの時間が経過する場合に、ACKが最終的に送信されることを確保するために使用される。ノードがメッセージを受信するとき、ノードは自身の遅延ACKタイマを始動させる。ノードがメッセージに確認応答するよりも前に遅延ACKタイマが終了する場合には、ノードはACKタイマ終了に応答してメッセージに確認応答することとなる。

【0038】図15のステップ1において、ノードA内の最終メッセージ宛先レジスタがセットされ、次のメッセージ(m1)が、最終メッセージの保管されている宛先と比較される。この例に関して、最終メッセージ(m1より前の)はグループXに対するものであったと仮定される。最終メッセージ宛先は、次のメッセージm2に対する宛先と同じであるので、遅延ACKフラグがセットされる。メッセージ・タイマが始動され(ステップ2)、m1が送信される(ステップ3)。図14を再び参照すると、この例に関して、LAN2(1410)に対するseq1(1420)およびseq2(1430)は、いずれも1であり、LAN3(1440)に対するseq1(1450)およびseq2(1460)は、いずれも50である。これらの数字、1および50は、任意であり、異なる値を割当ててLAN2とLAN3とに対するシーケンス番号を区別することを可能にする。

【0039】m1を発信することは、m1ヘッダ(図11を参照)内にseq1=1およびseq2=1を備え、1(真)にセットされた遅延ACKフラグを備えて

ノードBに対してm1を送信することによって行われる。メッセージm1は、続いて、同じ方法でノードDに対して送信される。メッセージm1は、続いて、seq1=50およびseq2=50を備え、1にセットされた遅延ACKフラグを備えてノードCに対して送信される。ノードB、CおよびDの各々がm1を受信するとき、それらは、それら各自の遅延ACKタイマを始動させ(ステップ1')、それら各自のCLUEにm1を渡す(ステップ2')。同じ値であるシーケンス番号seq1およびseq2の双方を有する全てのメッセージは、これが新しい送信ウィンドウの最初のメッセージであるという信号を受信側ノードへ送り、それが、全ての先のメッセージが送信され、未処理のメッセージが1つもなく確認応答されたということを指示するということ留意されたい。

【0040】次に、ノードAは、次のメッセージm3の宛先を検査し、それが現行メッセージm2の宛先と一致するか否かを確かめる。メッセージm3は、ノードAおよびノードB間の二地点間メッセージであり、一方、メッセージm2は、グループX内の全てのノード、すなわちB、CおよびDに対するものである。これらのメッセージの宛先は一致しないので、遅延ACKフラグはゼロ(偽)にセットされ、m2が送信される。ノードBおよびDに対するメッセージm2は、seq2を2にインクリメントし、スライディング送信ウィンドウが2つのメッセージ、m1およびm2に増加したということを指示する。一方、遅延ACKフラグは、メッセージm2においてクリアされ、それは、ノードB、CおよびDに、次のメッセージを発信するよりも前にスライディング送信ウィンドウ内の全てのメッセージに確認応答することを要求するということ留意されたい。ノードB、CおよびDは、m2に対する遅延ACKフラグがクリアされていることを確かめ、これは、それら各自に未処理のメッセージに確認応答することを要求する。まず、遅延ACKタイマがクリアされ(ステップ3')、m2が各自のCLUEに渡され(ステップ4')、どのメッセージが確認応答されているのかを指示するACKメッセージのシーケンス番号を用いることによってm1およびm2の双方を確認する単一のACKメッセージが、ノードB、CおよびDの各々からリターンされる。このようにして、ノードBおよびDはseq1=1およびseq2=2を用いて確認応答し、一方ノードCは、seq1=50およびseq2=51を用いて確認応答する。全ての未処理のメッセージに対するACKが受信されたので、この時点で、ノードAはメッセージの送信を再開することができる。

【0041】ノードAは、次に、その送信キューを検査し、m3がこの特定の時点において送信キュー内の最終メッセージであるということを調べる(ステップ6)。これに回答して、ノードAは、自身のメッセージ・タイ

10

20

30

40

50

マをリスタートし、メッセージ宛先をリセットする（ステップ7）。メッセージm3が続いて発信される。m3は、ノードAからノードBへの二地点間メッセージであるということを留意されたい。シーケンス番号は、3にインクリメントされ、遅延ACKフラグが真にセットされ、m3が続いて送信される（ステップ8）。これに回答して、ノードBは自身の遅延ACKタイマを始動させ（ステップ6'）、ノードBは自身のCLUEにm3を渡す（ステップ7'）。

【0042】続いて、メッセージm4が送信キュー内に到達すると仮定する。メッセージm4の宛先（グループX）は、最終メッセージm3の宛先（ノードB）と一致せず（ステップ9）、したがって続行するよりも前に、ノードAからノードBへの先の二地点間メッセージは確認応答される必要がある。ACK要求メッセージは、シーケンス番号seq1およびseq2をメッセージあるいは確認応答されるべきメッセージの値にセットし、ヌル・メッセージ・フラグを真にセットすることによって送信される（ステップ10）。ノードは、ヌル・メッセージ・フラグを未処理のメッセージに即時に確認応答すべきコマンドと解釈する。その結果、ノードBは自身の遅延ACKタイマをリセットし（ステップ8'）、ノードBはm3に対する要求ACKを渡す（ステップ9'）。好適な実施の形態は、ノードに全てのメッセージを受信して実行することを強制することとなり、ネットワーク帯域幅とCPUリソースとを使用する全てのメッセージの受信を、全てのノードが要求されるわけではないという点で特有である。かわりに、メッセージは、宛先が変わるときに保留メッセージに対するACKを強制するヌル・メッセージを用いて、自身の意図された受身側へのみ送信される。

【0043】この時点において、メッセージ・タイマがリスタートされ、メッセージ宛先がリセットされる（ステップ11）。続いてメッセージm4が発信される（ステップ12）。ノードBおよびDに対するシーケンス番号はいずれも4であり、一方ノードCに対するシーケンス番号は52であるということを留意されたい。ノードBは、これまでのメッセージの全てを確認しているの
40で、ノードBは、次のメッセージのシーケンス番号が4であると予期する。一方、ノードDは、ノードAからノードBへの二地点間メッセージm3を確認しなかったの
50で、ノードDは、次のメッセージのシーケンス番号が3であると予期することに留意されたい。好適な実施の形態に従うシステムのアーキテクチャは、予期された番号よりも大きいシーケンス番号を受信するノードは、送信側が全ての必要なACKを確認し、ノードは、それが受信しなかったメッセージを受信するよう要求されていなかったということを指示するseq1=seq2をセットした送信側を単に信用するように定義される。したがって、ノードDが4のシーケンス番号seq1およびs

eq2を有するメッセージm4を確認するとき、ノードDは、先のメッセージはノードDに向けられていなかったということを信用し、したがってm3の欠落を気にせずm4の処理に移る。

【0044】ノードAは、送信すべきメッセージをそれ以上有さないと仮定され、したがって、図15に示されるように、最終的にノードB、CおよびD上の遅延ACKタイマは、全て終了し（ステップ10'）、これらのノードの各々に、適切なシーケンス番号を用いてm4に対するACKを渡させることとなる（ステップ11'）。

続いてメッセージ・タイマがリセットされ、メッセージ宛先がリセットされ（ステップ13）、続いて、ノードAは次のメッセージを待つ。

【0045】図13～15の具体例は、実施例の具体的な詳細を含む。この実施例は、好適な実施の形態の顕著な特色の一部を説明するために示されており、限定と解釈されるべきではない。好適な実施の形態は、順序付きメッセージを使用するクラスタ・コンピュータ環境におけるスライディング送信ウィンドウを提供するあらゆるメカニズムおよび方法に明らかに及ぶ。

【0046】好適な実施の形態と関連して説明される本発明は、先行技術よりも著しく優れた改良をここに提供する。これにより、スライディング送信ウィンドウを、マルチキャスト・メッセージ可能なクラスタ・コンピュータ環境において用いることができ、これは、以前は行うことができなかった。スライディング送信ウィンドウを与えることにより、送信側は、次のメッセージを送信するよりも前にメッセージに対するACKを待つ必要がない。さらに、受信側は遅れて単一の確認メッセージを用いていくつかのメッセージの確認を送信側へ返すことができる。このように、クラスタはよりビジーなメッセージ処理を維持し、ネットワーク・トラフィックは著しく削減され、それによってシステムの性能が高められる。

【0047】当業者は、本発明の範囲内で多くの変形が実行可能であるということを理解できる。それゆえ本発明の好適な実施の形態と関連して本発明を詳細に示し説明してきたが、当業者は、本発明の趣旨および範囲から外れることなしに、形態および詳細についてのこれらのおよび他の変更をその中に行うことが可能であるということを理解できる。

【0048】まとめとして、本発明の構成に関して以下の事項を開示する。

(1) 装置であって、少なくとも1つのプロセッサ、前記少なくとも1つのプロセッサに接続されたメモリ、少なくとも1つの他のコンピュータ・システムに接続されたネットワークへ前記装置を接続するネットワーク・インターフェース、前記メモリに存在し前記少なくとも1つのプロセッサによって実行されるクラスタ通信機構、前記クラスタ通信機構は、次の順序付きメッセージを発

信するよりも前に少なくとも1つの他のコンピュータ・システムからの確認メッセージを待つことなしに、前記少なくとも1つの他のコンピュータ・システムへ少なくとも1つの順序付きメッセージを伝達するスライディング送信ウィンドウを有する、を備える装置。

(2) 各順序付きメッセージは、前記順序付きメッセージに対する確認メッセージを遅らせて少なくとも1つの後続の確認メッセージと共にグループ化できるか否か指示する情報を有するヘッダを含む上記(1)に記載の装置。

(3) 前記確認メッセージは、1から複数の順序付きメッセージを確認する上記(2)に記載の装置。

(4) コンピュータ・システムのクラスタを備えるネットワーク・コンピュータ・システムであって、前記コンピュータ・システムの各々は、各コンピュータ・システムを、前記クラスタ内の他のコンピュータ・システムへネットワークを介して接続するネットワーク・インターフェース、メモリ、前記メモリに存在するクラスタ通信機構、前記クラスタ通信機構は、次の順序付きメッセージを発信するよりも前に少なくとも1つの他のコンピュータ・システムからの確認応答を待つことなしに、前記少なくとも1つの他のコンピュータ・システムへ少なくとも1つの順序付きメッセージを伝達するスライディング送信ウィンドウを有する、を含むネットワーク・コンピュータ・システム。

(5) 各順序付きメッセージは、前記順序付きメッセージに対する確認メッセージを遅らせて少なくとも1つの後続の確認メッセージと共にグループ化できるか否か指示する情報を有するヘッダを含む上記(4)に記載のネットワーク・コンピュータ・システム。

(6) クラスタ・コンピュータ環境においてタスクを処理するコンピュータ実装方法であって、前記方法は、次の順序付きメッセージを発信するよりも前に、順序付きメッセージを受信した前記クラスタ内の各コンピュータ・システムからの確認応答を待つことなしに前記クラスタ内の少なくとも1つの他のコンピュータ・システムへ少なくとも1つの順序付きメッセージを伝達するスライディング送信ウィンドウを有する、前記クラスタ内の第1のコンピュータ・システム上で実行するクラスタ通信機構を与えるステップと、前記クラスタ通信機構が、前記クラスタ内の少なくとも1つの他のコンピュータ・システムへ第1の順序付きメッセージを送信するステップと、前記クラスタ通信機構が、前記クラスタ内の前記少なくとも1つの他のコンピュータ・システムからの前記第1の順序付きメッセージに対する応答を待つことなしに、第2の順序付きメッセージを送信するステップとを含む方法。

(7) 前記クラスタ内の前記少なくとも1つの他のコンピュータ・システムが、前記第1および第2の順序付きメッセージの双方を確認する単一の確認メッセージを前

記クラスタ通信機構へ送信することによって前記第1および第2の順序付きメッセージに応答するステップをさらに含む上記(6)に記載の方法。

(8) 前記第1および第2の順序付きメッセージは、各々、前記第1および第2の順序付きメッセージに対する確認メッセージを遅らせて少なくとも1つの後続の確認メッセージと共にグループ化できるか否か指示する情報を有するヘッダを含む上記(6)に記載の方法。

(9) プログラム・プロダクトであって、(A) 次の順序付きメッセージを発信するよりも前に、少なくとも1つの他のコンピュータ・システムからの確認応答を待つことなしにクラスタ内の少なくとも1つの他のコンピュータ・システムへ少なくとも1つの順序付きメッセージを伝達するスライディング送信ウィンドウを有するクラスタ通信機構(A1)を含むコンピュータ・プログラム、(B) 前記コンピュータ・プログラムを伝達するコンピュータ読み取り可能信号伝達媒体、を備えるプログラム・プロダクト。

(10) 前記信号伝達媒体は、記録可能媒体よりなる上記(9)に記載のプログラム・プロダクト。

(11) 前記信号伝達媒体は、伝送媒体よりなる上記(9)に記載のプログラム・プロダクト。

(12) 各順序付きメッセージは、前記順序付きメッセージに対する確認メッセージを遅らせて少なくとも1つの後続の確認メッセージと共にグループ化できるか否か指示する情報を有するヘッダを含む上記(9)に記載のプログラム・プロダクト。

【図面の簡単な説明】

【図1】ネットワーク上で相互通信できるコンピュータ・システムのブロック図である。

【図2】クラスタ・コンピュータ環境におけるマルチキャスト通信をサポートする先行技術ノード上で実行されるプログラムのブロック図である。

【図3】図2に示されるクラスタ通信機構260のブロック図である。

【図4】好適な実施の形態に従うクラスタ通信機構のブロック図である。

【図5】好適な実施の形態のスライディング送信ウィンドウの背景にある概念を示す図である。

【図6】クラスタ内のノードとして使用できる好適な実施の形態に従うコンピュータ・システムのブロック図である。

【図7】クラスタ内でローカル・エリア・ネットワーク(LAN)によって相互接続される3つの異なるコンピュータ・システムを示すブロック図である。

【図8】図7のノード間の先行技術対話を示す図である。

【図9】好適な実施の形態に従う図7のノード間の対話を示す図である。

【図10】好適な実施の形態に従ってスライディング送

21

信ウィンドウを実装する方法のフロー図である。

【図11】好適な実施の形態に従ってメッセージ・ヘッダに含まれる情報を示すブロック図である。

【図12】遅延ACKフラグが図12におけるヘッダのフラグ部分1114の一部であることを示すブロック図である。

【図13】異なるローカル・エリア・ネットワーク（LAN）上に位置する4つのノード間のハイブリッド・ネットワーク接続例を示すブロック図である。

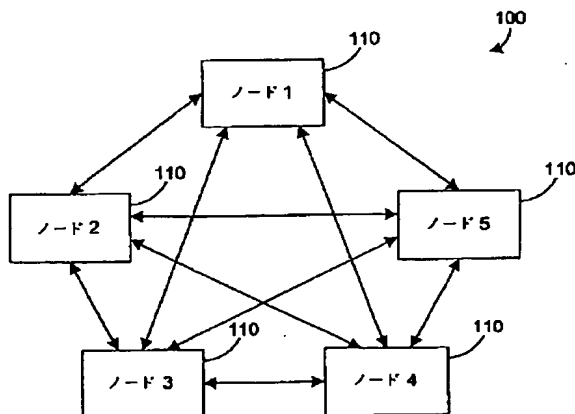
【図14】図13および15に示されるノードAの機能を示すブロック図である。

【図15】好適な実施の形態に従う図13のノード間の対話を示す図である。

【符号の説明】

100 クラスタ
110, 210 ノード
212 クラスタ・マネージャ
214 ジョブ/アプリケーション
220 クラスタ化APIコンポーネント
230 クラスタ・ライブラリ関数
240 クラスタ制御層
242 クラスタ・グループ・マネージャ
250 クラスタ・エンジン
260, 460 クラスタ通信機構
262 クラスタ・トポロジ・サービス
270 UDP層
280 IP/物理層
290, 690 システム・レベル・コード
292, 692 カーネル
310 固定送信ウィンドウ
320 送信キュー
330 現行メッセージ属性
340 ACKインジケータ

【図1】



22

* 410 スライディング送信ウィンドウ

430 現行メッセージ・キュー

440 保留ACKキュー

600, 675 コンピュータ・システム

600A, 600B, 600C ノード

610 プロセッサ

620 主記憶装置

622 データ

624 オペレーティング・システム

10 626 IPマルチキャスト・サポート

630 大容量記憶インターフェース

640 端末インターフェース

650 ネットワーク・インターフェース

655 DASD

660 システム・バス

665 端末装置

670 ネットワーク

695 フロッピー（R）・ディスク

700 ネットワーク構成例

20 1100 ヘッダ

1110 バージョン・フィールド

1112 タイプ・フィールド

1114 フラグ・フィールド

1116 長さフィールド

1120 ソースIDフィールド

1130 宛先IDフィールド

1140 ソースIPフィールド

1150 宛先IPフィールド

1160 コネクション番号フィールド

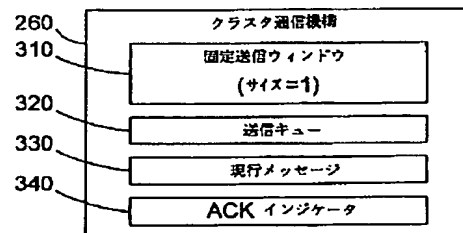
30 1170, 1180 シーケンス番号フィールド

1190 ネクスト・フィールド

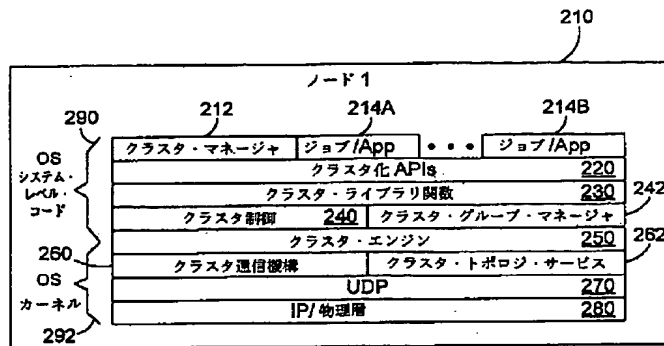
1200 遅延ACKフラグ

* 1410, 1440 オブジェクト

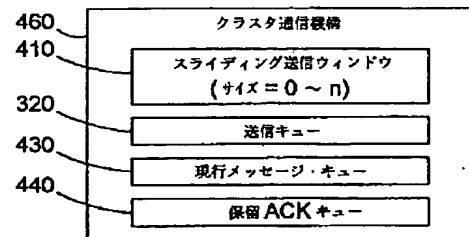
【図3】



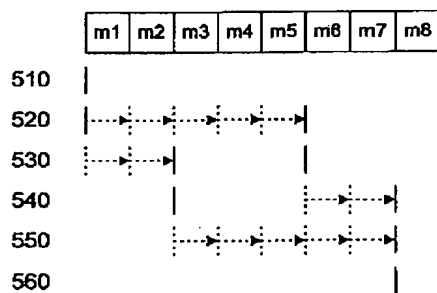
【図2】



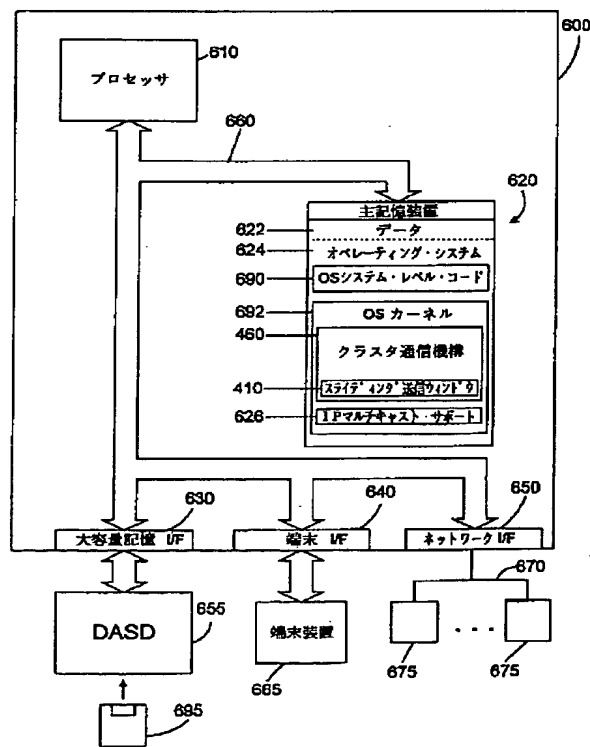
【図4】



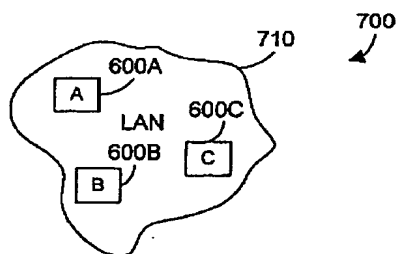
【図5】



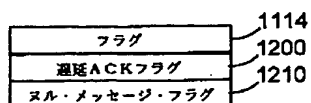
【図6】



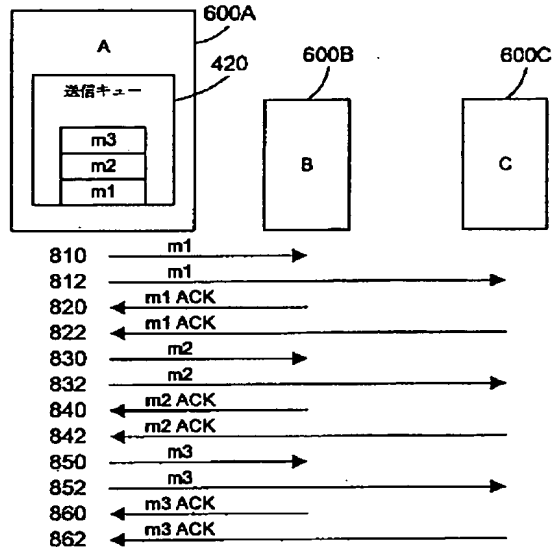
【図7】



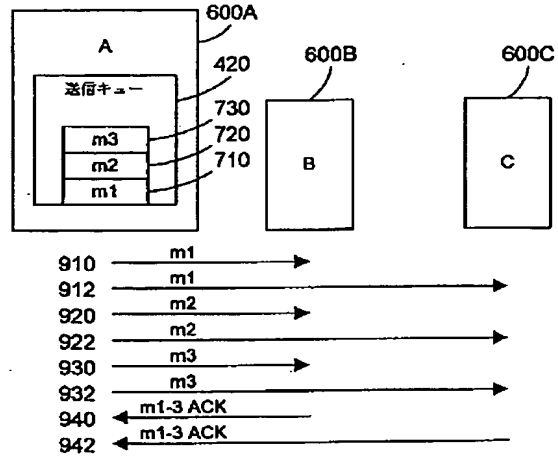
【図12】



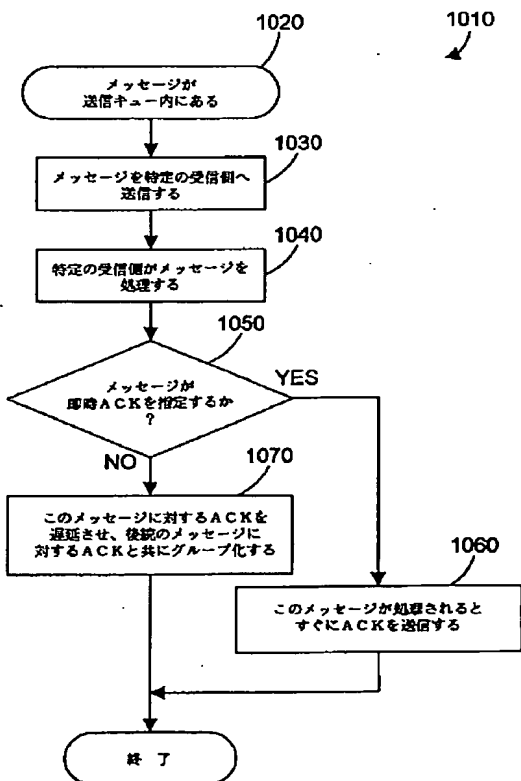
【図8】



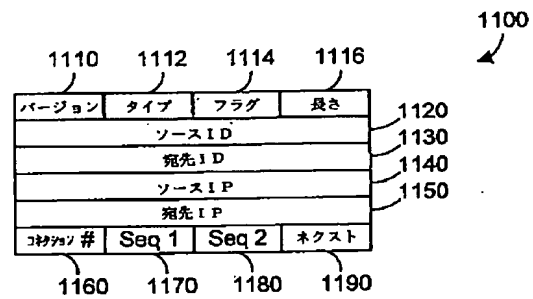
【図9】



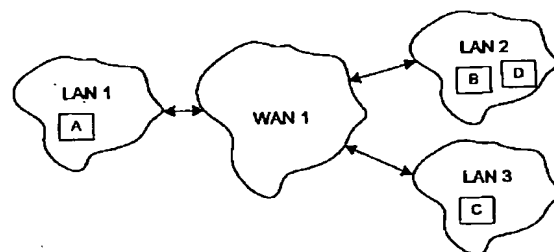
【図10】



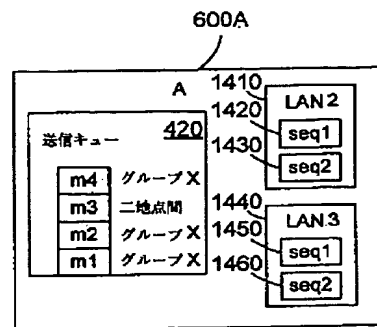
【図11】



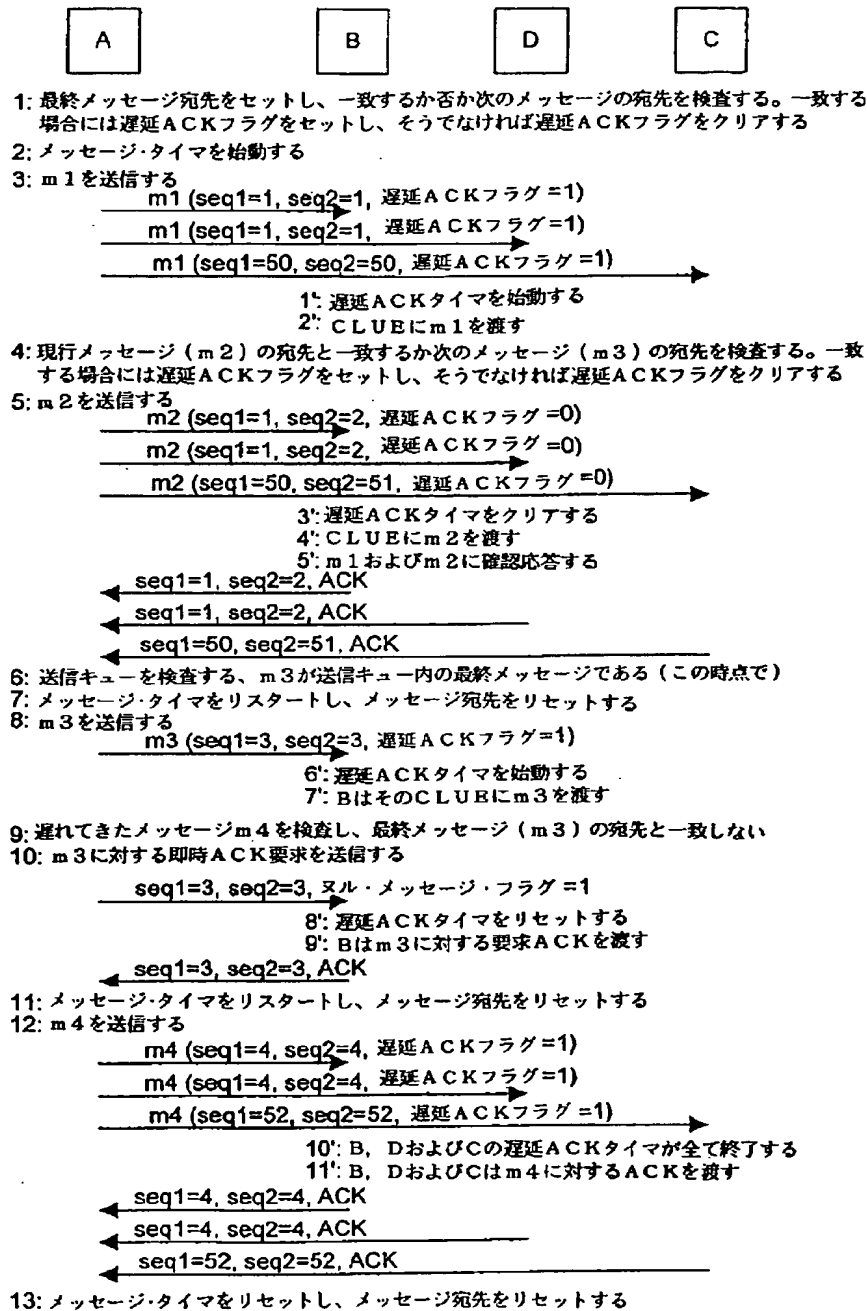
【図13】



【図14】



【図15】



フロントページの続き

(72)発明者 ティモシー・ロイ・ブロック
アメリカ合衆国 55901 ミネソタ州 ロ
チェスター エイヴォン レーン エヌダ
ブリュ 4516

F ターム(参考) 5B045 BB56 GG01
5B098 AA10 GA04 GD02 GD14
5K033 AA01 AA02 BA04 CA06 CB01
CB04 CB06 CB13 CC01 DB16
5K034 AA02 AA07 EE10 HH01 HH02
HH08 MM18